Exploring the Use of Limited Proteolysis Mass Spectroscopy to Investigate Protein Structure and its Potential Application as a Screening Tool

Michael Todd Brown

Molecular Biology

Dr. Jennifer Garcia

Abstract

A protein's function is highly informed by its structure. Understanding the structural nuances of proteins could greatly inform studies into their function. Limited proteolysis-Mass spectrometry (LiP-MS) is a novel method for observing the structure of proteins. LiP-MS determines protein structures through treating complex protein samples (such as blood) with a broad specificity protease for a small amount of time to cleave accessible regions of the peptide backbone. Then the protein sample is fully digested with sequence-specific proteases to make fragments suitable for MS analysis. This treatment allows LiP-MS to determine global protein structural changes such as those mediated by changes when exposed to environmental stimulus. Next, LiP-MS allows for a targeted analysis of protein structure in complex solutions. However, there are limitations of LiP-MS. The visualization of proteins is only as informative as the variety and breadth of LiP-MS ratios generated. As of now the range of LiP-MS ratios does not provide a clear enough picture for medical use. For our study, we aim to optimize the LiP-MS protocol to better its accuracy in quantifying the presence of misfolded proteins found in cerebrospinal fluid of individuals with Alzheimer's disease. With this optimization of the LiP-MS protocol, we hope to investigate proteins associated with Alzheimer's disease in their misfolded state to create a screening assay for earlier diagnosis and prognosis of Alzheimer's disease and other neurodegenerative diseases. Prior to our study, LiP-MS was unable to provide detailed information on the structural changes that occurred. We altered a variety of experimental LiP-MS variables and compared them to current, standard LiP-MS protocols. We found that though some aspects could change, the current experimental variables are effective at generating LiP-MS ratios. Further work is needed to improve LiP-MS to potentially allow for screening against diseases that involve protein misfolding, such as Alzheimer's disease.

Introduction

Protein Structure, Structural Alterations, and Protein Functions

Life relies on complex interactions between proteins and macromolecules. Proteins act as the biological tools for cellular processes with varying roles. Due to their role in various biological processes, protein structure is highly dynamic and modular, and a protein's purpose in the cell is dictated by its structure. For proteins to adopt functional structures, they must fold together from their initial polypeptide string. The folding process is the result of complex chemical interactions between the 22 amino acids, the environment in which they exist, and energy dynamics that facilitate strings of peptides to fold into a three-dimensional structure. The final structure of proteins is dictated by which amino acids are used and the order of the amino acids. The complexity of the folding process and errors associated with folding can be consequential for the function of a protein. For example, mutations in DNA or errors in messenger RNA (mRNA) translation may lead to the content or order of aminos in the polypeptide string to differ, forming a variant that can misfold and thus lose critical function. Whether the errors occur in transcription or translation, the resulting error prevents the protein from correctly folding after production. When proteins fail to adopt their final structure, they are labeled as misfolded proteins. Misfolding can lead to a variety of outcomes for the protein's activity and presence in the cell.

Diseases of misfolded proteins

Often, the original function of a misfolded protein is lost, and the consequence of losing the protein's function can be detrimental to the cell. Due to the interconnected nature of biological functions and the vast network of interactions present, the loss of functionality could lead to decreased cell viability and can be the underlying cause for many diseases. For example, a loss of function mutation, in the human amyloid precursor protein (APP) can lead to the dangerous buildup of amyloid deposits that are strongly correlated with Alzheimer's disease. Another example is the misfolding of the human protein, Alpha-Synuclein, which is a major precursor in the development of Lewy bodies that lead to Parkinson's disease (Bross et al., 1999).

Understanding amyloid fibril formation is of great medical interest as there is a strong relationship between spontaneous protein aggregation and severely debilitating diseases like Alzheimer's disease, type 2 diabetes, prion diseases, Parkinson's disease, senile systemic amyloidosis, and Huntington's disease (Fontana et al., 2004). The common thread among these conditions is the presence of amyloid fibrils that can cause plaque buildup that negatively affects the health of the cells, eventually leading to damaged organs and decreasing health in an individual. A fundamental question to understand the pathology of these diseases is to determine the mechanism by which amyloid fibrils form from misfolded proteins and the exact nature of structural anomaly of amyloid fibrils. Two significant observations have been identified to cause a build-up of the protein aggregates into harmful structures: the existence of the cross-Beta structural motif and an unusual resistance to proteolytic degradation (Fontana et al., 2004). In both cases, neither observation is heavily understood, but more accurate visualization technology could unravel why these structural motifs and characteristics are present in a large number of harmful amyloid fibrils. By understanding the mechanism of amyloid fibril stability and their formation in the body, therapeutic strategies and drugs could be developed to combat the build-up of harmful protein aggregates and treat amyloid diseases.

Protein aggregates in neurological diseases

Protein structure and structural alterations are intimately linked to the mechanisms of amyloidosis, the aggregation of amyloid fibrils in organs that will eventually lead to organ damage and failure. According to recent studies, amyloid fibril formation from native proteins occurs via a conformational change in the proteins that lead to partly folded intermediates forming prefibrillar species that will further associate into mature fibrils (Fontana et al., 2004). The formation of intermediates is necessary for the formation of mature fibrils. Thus, one avenue of study lies in understanding the initial conformational changes that occur in early disease progression. This line of study could offer insights into how to monitor and prevent the progression of the disease. Not only is the initial conformational change important, but the subsequent structures that proceed the mature fibrils also offer avenues for treatment and medicine. In this research, we focus on nerve growth factor, VGF. VGF is a neurosecretory

protein involved in neurogenesis and neuroplasticity. VGF is also associated with learning, memory, depression, and chronic pain. Structural alterations to VGF that limit or compromise its function can lead to neurodegenerative diseases, like Alzheimer's disease through the build-up of misfolded proteins into large aggregates.

Methods are being developed to screen for misfolded VGF before they become mature amyloid deposits. This requires an efficient and quick method to visualize the protein structure and its structure alterations. Further, to make a significant clinical impact, this method would need to accurately identify misfolded variants that will lead to protein aggregation. With technology that can accurately and effectively provide protein structural information, the potential to develop a preventative medical screen that could identify neurodegenerative diseases in early states becomes more realistic. For example, if we could visualize or identify misfolded proteins variants or fragments before they eventually create large amyloid deposits, we may be able to administer preemptive treatment to Alzheimer's disease and lessen or prevent the onset of this debilitating disease.

Limited proteolysis-MS (LiP-MS) and Proteases

One technique that has shown potential towards a clinical method to study structural changes associated with disease, called Limited Proteolysis-MS (LiP-MS), involves protein fragmentation using protease treatment. While this method still needs investment to reach a significant medical impact; a major benefit of utilizing proteases lies in their ability to provide information on protein surface topology and mobility when combined with mass spectroscopy (MS). When paired with MS, proteases can offer a dynamic picture of the proteins structure and identify areas that are sensitive to structural changes caused by protein misfolding (Monti et al., 2005).

Proteases have several functions. For example, proteases regulate the fate, localization, and activity of many proteins, modulate protein-protein interactions, create new bioactive molecules through proteolytic cleavage, contribute to cellular information processing, and generate, transduce, and amplify cell signals. As a result, proteases influence multiple aspects of biology. Proteases have utility in research to understand and visualize protein folding.

Researchers use specific proteases to cleave proteins into peptides for MS analysis. A mass spectrometer ionizes peptides into a gas and filtered into a vacuum. This ionized peptide can then be analyzed to determine its mass and charge with a detector. This information can then be analyzed computationally to determine that peptide's sequence and abundance in complex samples (Glish & Vachet, 2003).

Experimentally, several proteases are used to understand how a protein can fold. Some proteases, called specific proteases, have specificity in where they cut polypeptides. These proteases cleave at a specific amino acid sequence leading to the production of peptide fragments after digestion. When digesting a protein with trypsin for mass spectroscopy, assuming you know the peptide sequence, a predictable set of peptide fragments will be generated in solution. The abundance of those peptide fragments can be measured using mass spectroscopy, thus allowing for quantification the presence of a specific protein or a set of proteins in a complex solution. This information can tell us to what degree a specific protein, or set of proteins' are present or absent, this, allows you to correlate protein presence and abundance to a specific phenotype. We can make specific claims about the abundance of a protein or a set of proteins in a complex solution like blood. (Kaur et al., 2018)

However, on its own mass-spectrometry with a standard trypsin digest cannot give specific information about protein structure. But when non-specific proteases are also used with mass-spectrometry, more information can be determined about proteins in complex solutions. The difference between specific and non-specific proteases revolves around their specificity. A specific protease, like trypsin, will always cleave at lysine and arginine at the carboxyl end. Non-specific proteases, like proteinase K, will cleave indiscriminately. Because of this, non-specific proteases will often cleave exposed regions of the protein first, as these regions of the protein are easier to access than a region embedded in the core of the protein. So, when a non-specific protease is paired with standard digestion done for mass-spectrometry, the non-specific protease can add another dimension of analysis to give information on accessible regions of a protein. A short treatment with a non-specific protease can give

insight into the surface topology of proteins or allow you to detect regions of proteins that undergo conformational changes in complex solutions. (Glish & Vachet, 2003)

The combination of standard trypsin digest and non-specific protease digestion is utilized in a technique called limited proteolysis (LiP). Limited proteolysis utilizes the predictable peptide generation of the standard trypsin digest while using protease accessibility visualization of non-specific proteases to generate structural data on a large variety of proteins in complex solutions like blood or cerebrospinal fluid. By measuring protease accessibility, Limited proteolysis can visualize changes to the surface topography of proteins that would indicate protein misfolding. Limited proteolysis also allows you to analyze a number of proteins in complex samples very quickly, meaning that there may be even more research applications beyond developing a screening method for misfolded proteins and be applicable beyond early detection of neurodegenerative diseases.(Mateus et al., 2021)

Limited proteolysis mass spectroscopy and screening for Alzheimer's Disease

The research described in this thesis revolves around exploring limited proteolysis-based mass spectroscopy's potential as a screening device for genetic and neurodegenerative disease. Because some genetic or neurodegenerative diseases begin as misfolded proteins that affect cell viability, LiP-MS is perfect for assessing the structure of specific proteins, such as VGF, to discover potentially misfolded proteins in patient samples. LiP-MS can also be done in complex matrices like blood or cerebrospinal fluid, allowing for samples from patients to be analyzed directly without purifying proteins from a sample. More importantly, LiP-MS would detect potentially misfolded proteins such as those that form prior to aggregate formation that are drug and protease resistant. Potentially identifying treatable misfolded proteins can allow for more effective treatment before symptoms manifest. However, for LiP-MS to be an effective screening tool, it requires optimization to produce reliable and reproducible results in a clinical setting. For this thesis, LiP-MS was performed using various experimental conditions to increase the range of readings that describe the accessibility of regions of proteins. The quantitative assessment of protease accessibility is called a LiP ratio. LiP ratios are created by dividing the relative intensity of a peptide fragment in a LiP aliquot (where the sample is treated with both a non-specific and

specific protease) by the relative intensity of a corresponding peptide fragment in a control aliquot (where the sample is treated with just a specific protease). The resulting ratio usually exist on a scale from 0 to 1 with 0 indicating high protease accessibility and 1 indicating low protease accessibility.

Material and Methods

Prepping and Troubleshooting the Mass Spectrometer for Samples

An Orbitrap Fusion EcLiP-MSse mass spectrometer was used to analyze all of our cerebrospinal fluid samples. To operate the machine, it was necessary to manufacture all of the individual parts of the liquid chromatography setup. Columns were created by pulling 40 cm of 75 mm diameter silica piping taunt and precisely heating the center with a laser. The columns were then packed using pressurized gas to force silica packing beads through the column. Approximately 15 cms of the column were packed tightly. Traps were also manufactured in the lab and 15 cms of 150 mm diameter silica piping. Casein is used to seal one of the ends of the trap. Silica packing material is fed through the trap using pressurized gas. Approximately 4 cms of the trap were packed with silica beads. The LC apparatus must be constructed outside the MS.

The LC apparatus and mass spectrometer were run through multiple troubleshooting programs to check for leaks throughout the instruments. The pumps were also primed and check to ensure that they were functioning properly.

A Peptide Retention Time Calibration Solution Was Used to Calibrate the Mass Spectrometer

Forty µl aliquots of cerebrospinal fluid (CSF) were created from a 1ml aliquot of CSF. The 40µl aliquot is split into two separate aliquots, one the control, the other LiP-MS. Both aliquots make up the one-off sample. Both aliquots in a sample received 2 µl of QconCAT stock to a final concentration of 443 ng per aliquot. One point five µl of yeast alcohol dehydrogenase is added to both aliquots in a sample to a final concentration of 750 ng per aliquot.

Both the temperature treatment sample and chemical treatment sample received alterations to their protocol in the form increased temperature or exposure to Sodium deoxycholate conditions before

continuing with the standard LiP-MS treatment. Both aliquots in the sample received each treatment. The temperature treatment aliquot was placed on a heat block set to 98 degrees Celsius for 5 minutes. The chemical treatment was treated with 20 µl of DOC and left to sit for 20 minutes before protease K treatment.

The aliquots in each sample were then placed on a heat block for 1 hour at 25C. Proteinase K (PK) is added to both the aliquots in a sample at a 1:100 ratio for every aliquot variation except the 1:25 dilution and 1:50 aliquot variation which used their respective dilution ratios. Both aliquots were incubated after PK treatment for 1 minute at 25C. Both aliquots were then transferred to a 98-ºC heat block for 5 minutes to inactivate the PK. Two µl of yeast enolase stock were added to each aliquot to a final concentration of 400 ng per aliquot. DOC is added to a final concentration of 5 % (wt/vol) using 20 µl of stock DOC per aliquot in the chemical modification sample. One µl of 500 Dithiothreitol (DTT) is added to each aliquot to a final concentration of 12 mM. Both aliquots were incubated for 30 mins at 37C. Three point three tree µl of 500 mM Indole-3-acetic acid (IAA) is added to each aliquot. Both aliquots were incubated for 45 mins at room temperature in the dark. 1 µl of 2 µg/µl LysC is added to 20 µl of water. The 20 µl LysC/water solutions were added to each aliquot with a final concentration of LysC for each aliquot being 1:100 dilution. The aliquots were incubated at 37 degrees celsius for 4 hours with 800 rpm of agitation. The sample is diluted with ammonium bicarbonate to a final concentration of 1% using a stock solution of 100 mM AmBic. Trypsin is added to each aliquot to a 1:100 dilution ratio and a final concentration of 0.2 µg/µl. Both aliquot tubes were incubated overnight at 37 degrees Celsius at 800 rpm agitation. The digestion is ended for both aliquots and the DOC is precipitated by adding formic acid 95% formic acid to a final concentration of 2%. The acidified digest in both aliquots was incubated for one hour at room temperature to facilitate precipitation. The sample aliquots were centrifuged at 16000g (12000 rpm) for 10 minutes at room temperature. The clear liquid present in the aliquot after centrifugation was transferred to a new microcentrifuge tube. The centrifugation step is repeated.

Peptide Clean-up

After performing our limited proteolysis digests and standard trypsin and LysC digests for our sample aliquots, each sample aliquot was run through peptide clean-up protocol. The MCX columns for peptide clean-up were conditioned for use. 1 ml of methanol followed by 1 ml of 10% NH4OH in water, 2 ml methanol, 3 ml 0.1 % formic acid in water were used to condition the column, in that order. The sample from both aliquots was added to the MCX column. The MCX column was washed off its salts using 1 ml of 0.1% formic acid. The MCX column was washed off its neutrals with 1 ml of 90% acetonitrile/10% water. The peptides were eluted from the column with 1 ml of 10% NH4OH in methanol into a new low-bind microcentrifuge tube. The samples were speed vacuumed at 25 degrees Celsius. The samples were resuspended in 0.1% (vol/vol) formic acid to a final concentration of ~1 mg/ml. 1 µl from every sample aliquot to create a chromatographic library.

Mass Spectroscopy Analysis and Data Independent Acquisition (DIA) Analysis.

The sample peptides were loaded into the LC. The samples were run in batches of 5 and 6. In between the second and third run, the chromatographic library aliquot was run. The mass spectrometer ran all aliquots successfully.

PanoramaWeb, Cromwell workflow and Skyline

Using a Cromwell workflow, the data from the mass spec analysis was converted into a format that could be viewed and altered in Skyline. The raw data files from the mass spectrometer are uploaded to PanoramaWeb. Then they are run through a workflow that converts the files into an MzML format. MzML files are compatible with skyline software. The chromatographic library files are created using EncyclopeDIA to create an elib file that is also viewable in Skyline. The data is then uploaded to Skyline and exported to an Excel sheet.

RStudio analysis and PyMOL

Using R studio, the mass spectroscopy data was analyzed to look at LiP-MS ratios for all the peptides in solution. An excel sheet containing all the data from mass spectroscopy analysis was exported from Skyline and imported into RStudio. The data was then parsed and cleaned for NA values and the

Total Ion Current Area category of Mass spectroscopy analysis was removed from the data set. The Replicate column is separated into 6 separate columns, specifically, into year, month and day, Instrument, project, Reference number, and aliquot columns. A column was then added to the data based on aliquot column that separates aliquots into their control and LiP-MS categories respectively. The LiP-MS ratio column was added by dividing all total area of the LiP-MS aliquots by the total area of their paired control aliquot. A column was added to describe the experimental conditions of each aliquot. Various graphs were created in RStudio to reflect how experimental conditions resulted in changes to LiP-MS ratios observed. The data was filtered to only include LiP-MS ratios from peptides originally from the VGF protein. Then the data for the VGF protein was further separated by aliquots. A new data-frame was created using a while loop to match LiP-MS ratios with the region of the protein they correspond to. This data-frame was converted into a text file. A python script was used in conjunction with the newly acquired text file to highlight various regions of the protein in PyMOL according to LiP-MS ratios with 0 to 1 being colored a gradient of blue and 1 to 3 being colored a gradient of red.

Results

Explanation and Optimization of LiP-MS Ratios

The purpose of our research was to test and optimize the experimental conditions in the standard limited proteolysis mass spectroscopy (LiP-MS) method to gather information about protein structure in a complex solution like cerebrospinal fluid samples. Standard protocol LiP-MS can visualize multiple areas of protease accessibility for all proteins in the proteome. There are currently limitations in the LiP-MS protocol that effect its ability to produce high fidelity and accurate determinations of protein structures in complex solutions that prevent the method from being useful in a medical context.

The current standard LiP-MS protocol uses a 1:100 dilution for Proteinase K and allows Proteinase K (PK) to digest the sample for 1 minute, before using fresh LysC to prime samples for

standard trypsin digest and mass spectroscopy analysis. We ran two different cerebrospinal fluid samples, each collected from a different patient through the standard protocol (Figure 1). The data we gathered from these samples has two major problems: consistency, and distribution. The distribution of LiP-MS data from Standard protocol is less uniform with clear peaks in the violin graph. This limits the ability to compare samples due to tight range of LiP ratios obtained. It is also largely apparent that the standard protocol does not produce consistent results from sample to sample as we see vastly different distribution of LiP ratios between the two samples used. This suggested that LIP-MS required alterations to the standard protocol to get more reproducible and informative data.

Ideally, LiP-MS optimization will provide an abundance of evenly distributed LiP-MS data that can tell us more about the protein structure. LiP-MS data allows for analysis of protein structure by numerically quantifying the accessibility of 4-40 amino acid long regions of the protein. The regions or windows of data correspond to the individual peptide fragments generated by trypsin digest, as such they range from 5 to 40 amino acids long.. A LiP-MS ratio is computed by dividing the LiP-MS aliquot intensity value for the peptide fragment of interest by the control aliquot intensity value for the peptide of interest. The reason that LiP-MS ratios can be computed relies on the lack of the peptide of interest in the LiP-MS aliquot due to extensive fragmentation by proteinase K. LiP-MS ratios can be interpreted as such: informative LiP-MS ratios exist from a range of 0 to 1. LiP-MS ratios close to zero indicate high protease accessibility as there are less peptide fragments in the LiP-MS aliquot as opposed to the control aliquot, thus moving the LiP-MS ratio closer to zero. LiP-MS ratios close to 1 indicate regions with low protease accessibility as there are the same number of peptides in the control and LiP-MS aliquot, indicating the non-specific protease was not able to access this region as it had less accessibility.

Secondly, ideal LiP-MS ratio data will contain an even distribution of ratios from 0 to 1 so that the number associated with a protein fragments can give a quantitative value of accessibility that can be used to give nuanced information about the structure of the particular protein. Data generated where LiP-MS ratios “clump together” (i.e. the majority of the LiP ratios share the same value) is less informative for the analysis of protein structure. The clumping of values eliminates meaningful differences between

how accessible one peptide sequences in a protein structure is to another. Ideally, we want to obtain a wide distribution of LiP-MS ratios between 0 and 1 to make accurate and meaningful determinations of accessibility and surface topography.

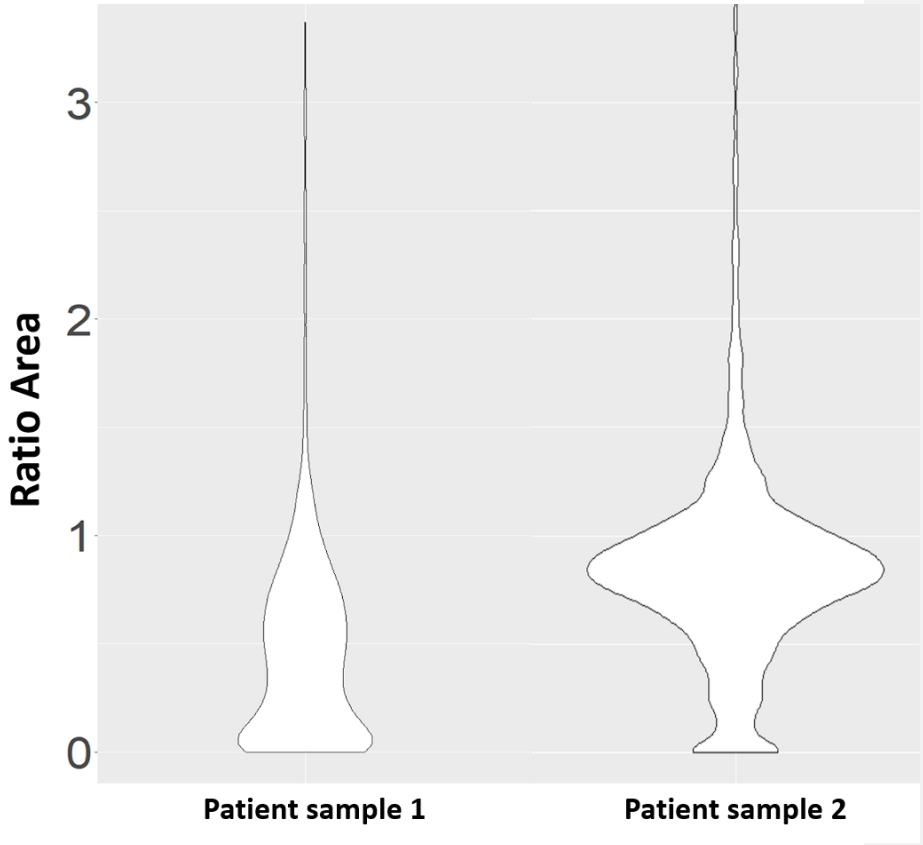

Figure 1. A comparison of two patient aliquots
The two samples used to test the effectiveness of LiP-MS and the various experimental changes are very different in their distribution of LiP ratio. This indicates that LiP-MS protocol needs to be optimized for medical use as there is no consistent observable patterns making comparison between medical treatments and experimental treatment groups difficult.

LiP-MS's inability to produce consistent results is present at the individual protein level as well. Using PyMOL computer software we can color code regions of a protein to correspond to the LiP ratio

experimentally obtained with LiP-MS. In Figure 2, blue coloring indicates a high protease accessibility for the non-specific protease PK and white indicates low protease accessibility for the non-specific protease PK. Red indicates that the LiP-MS ratio for that region exceeds 1. Figure 2 compares the distribution of LiP ratios on the protein VGF and the effects of trying to simulate misfolding using artificial means. Because we are trying to artificially misfold the protein we expect to see increase protease accessibility throughout the protein structure due to the disruption of the native structure and subsequent denaturing that follows. Instead, the two misfolding simulation aliquots, temperature treatment and sodium deoxycholate (DOC) treatment, show the same regions of accessibility as the control. Furthermore, these regions are not more accessible, rather some areas became less accessible specifically in the case of sodium DOC treatment where multiple regions that are accessible in the control become less accessible when treated. The temperature treatment shows some areas of increased accessibility yet suffers from a similar problem as there are regions that are accessible in the control that become less accessible with the treatment. The fact both treatment aliquots show highly unexpected and varied results from our hypothesis further supports the need to optimize the protocol to increase the efficiency and consistency of the results.

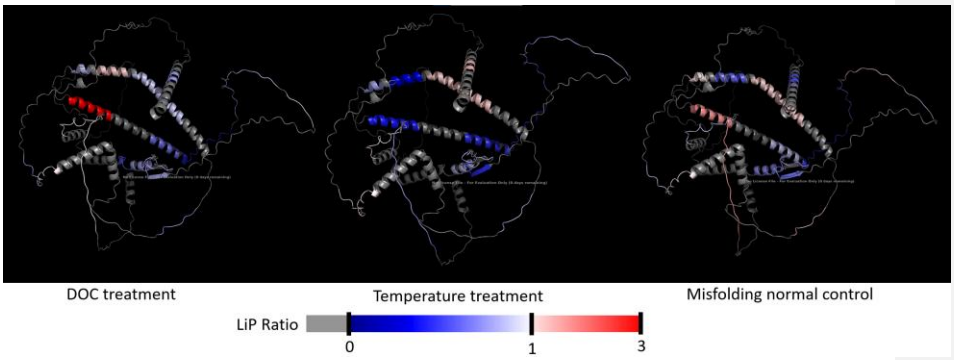

Figure 2. VGF PyMOLe structural diagrams displaying regions of protease accessibility, misfolding simulation sample group
Using PyMOLe, regions of proteins with higher and lower LiP ratio values have been color coded to reflect their value. Darker blue regions of the protein indicate regions of high protease accessibility, whereas red regions indicate regions of low protease accessibility. The protein in question is VGF. The sample group presented is the misfolding simulation group. The differences in protease susceptibly

illustrate that the temperature treatment was more effective in altering the structure of the protein and mimicking a misfolded protein.

We aimed to improve LiP ratios created from cerebrospinal fluid samples from volunteers. LiP ratios were gathered using the accepted standard protocol for Limited proteolysis. Our study sought to increase the range of LiP ratios compiled from these samples so that we can create more accurate and reliable patterns for misfolded variants. We tested a variety of experimental variables. We tested if varying the concentration or digestion time of Proteinase K, the non-specific protease, added to a LiP aliquot could alter the LiP ratio. We also tested the effect of endo-proteinase LysC affected LiP data quality. LysC is a protease that cleaves peptide bonds at the carboxyl side of lysine during digestion so that more regions can be accessed by the specific protease. We tested the effects of a new and old LysC enzyme as well as the removal of LysC from the protocol to see the overall effectiveness of LysC treatment in generating useful LiP-MS results. With refinement of the LiP-MS protocol, we hope that LiP-MS could be used to screen against disease before symptoms manifest. Our study was able to find multiple experimental conditions that can be altered to optimize LiP-MS so that it can be used as a medical screening technology.

However as apparent with the data obtained in Figure 5, 6 and 7, the LiP-MS protocol has the potential to generate values that exceed 1. Treatment of the cerebrospinal fluid samples using the standard LiP-MS protocol generated LiP-MS ratios with values up to 3. A LiP-MS ratio above one does not provide informative data on protease accessibility and does not give more detail to the surface topography of the protein. LiP-MS ratios above 1 are also not valuable for interpreting protein misfolding in complex samples. LiP-MS ratios that exceed 1 indicate that there was more of the peptide of interest in the LiP-MS aliquot vs. the control, or the opposite of what you would see for high proteases accessibility. There are two major ways to interpret LiP-MS ratio values that exceed 1. The first explanation rationalizes that LiP-MS ratios greater than 1 can indicate that a more mobile peptide sequence is being preferentially targeted by the longer digestion time in the LiP-MS aliquot versus the control. Mobile regions of protein are often

protease accessible, and these mobile regions are accessible by both proteinase K and trypsin. In the case that both treatments target a peptide sequence, it is possible that the LiP-MS aliquot will have more of the peptide fragments generated than the control because of an extended treatment with both kinds of proteases. This can lead to the abundance of the specific protein fragment in the LiP aliquot to exceed the number of peptide fragments created in the control aliquot, thus, leading to a ratio larger than 1 (Figure 3). The second explanation for LiP-MS ratios greater that 1 posits that non-specific proteases are able to cleave away regions of the proteins so that the standard digestion is able to produce more peptides fragments because of greater accessibility to previously “guarded” regions of the protein. The analogy for the second explanation would liken the LiP treatment to be the pruning of a highly dense bush to get to specific interior branches, with the non-specific protease treatments acting as the pruning step in limited proteolysis treatment to give better access to the buried regions of a polypeptide (Figure 3).

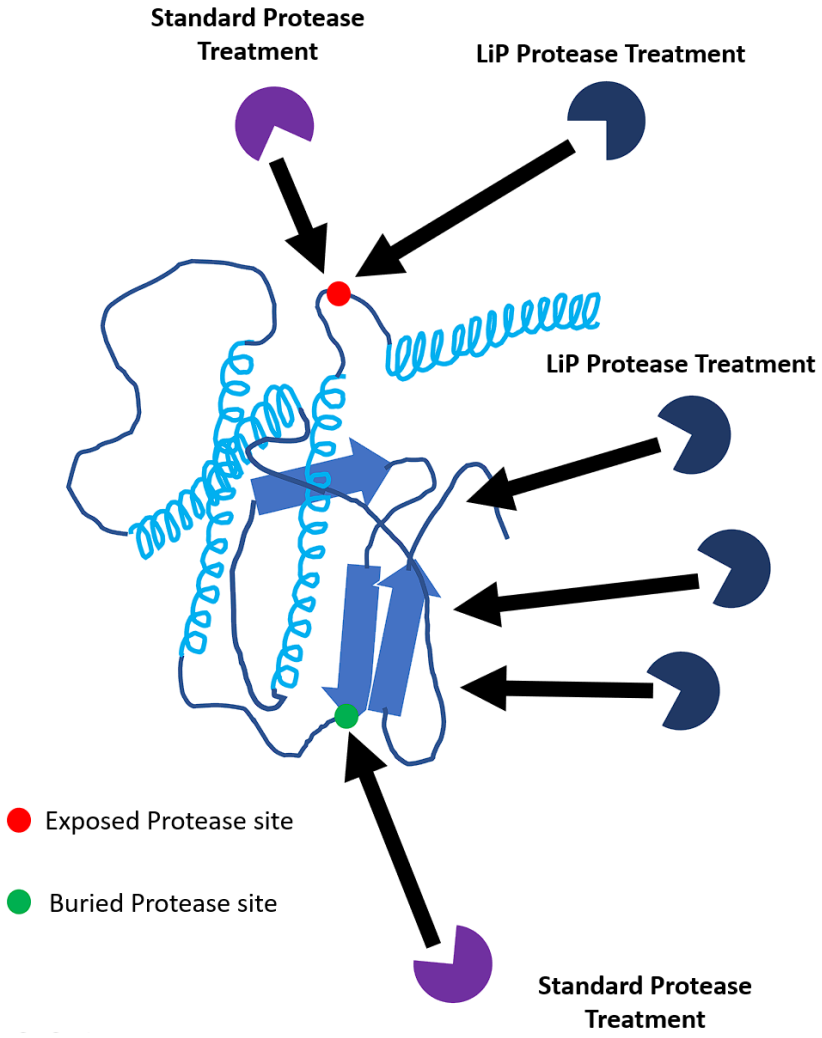

Figure 3. Highly mobile regions of a protein can be preferentially targeted by both non-specific and specific Proteases and buried regions can be opened up with initial nonspecific protease digestion. The highly mobile region of the protein is targeted by both the standard Protease treatment and the LiP-MS proteases treatments. The peptide is more abundant in the LiP-MS aliquot compared to the control aliquot and causes the LiP-MS ratio of the specific region to exceed 1.

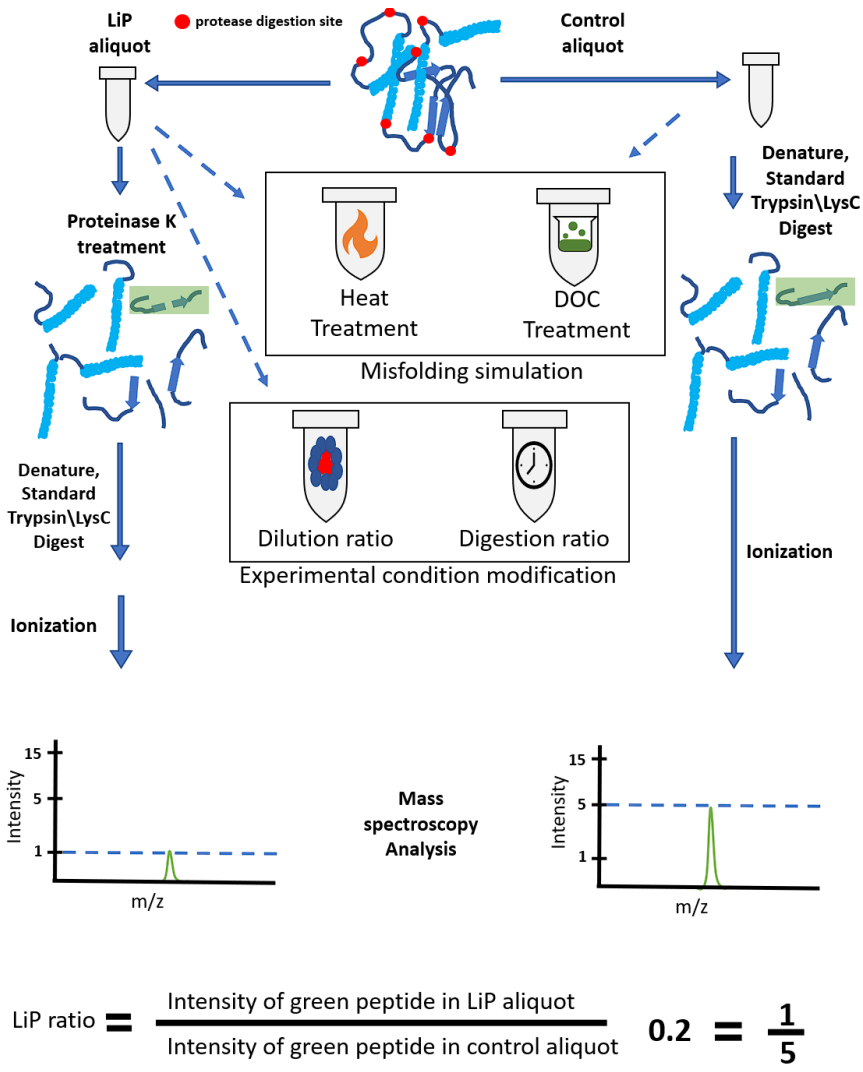

Figure 4. Limited proteolysis workflow.
The limited proteolysis workflow. During limited proteolysis, aliquots are separated into two groups. LiP-MS samples are treated with specialized reagents for limited proteolysis. The control aliquots lack the special Limited proteolysis treatment. The boxes labeled misfolding simulation and experimental condition modification indicate what portion of the workflow was alerted to produce specific experimental groups. The final portion of the figure illustrates how the mass spectrometer analysis differ between the LiP-MS and control aliquots.

Because we were not obtaining reproducible results between CBF samples, we next tested if the standard LIP-MS protocol can detect unfolded or denatured protein structures. To do this, we tested if the addition of sodium deoxycholate (DOC), a detergent that denatures proteins, or significantly raising the temperature of the sample before the addition of PK are able to affect the distribution of LiP ratios. Presumably, by stimulating misfolding by either DOC or heat, protein structures should become more accessible and shift the LiP-MS ratios towards 0. We compared LiP-MS ratios obtained from a DOC or heat-treated aliquot to a control aliquot that lacked either treatment. Both the LiP ratios obtained from the DOC treatment and temperature treatment have similar clumping of LiP-MS ratio values towards 1, where the misfolding simulations are slightly increased relative to the control (figure 6). The means for each aliquot are consistent with the observations of clumping in the violin graph. The mean for the DOC treatment is 1.06. The mean for the temperature treatment is 1.01. The mean for the misfolding control is 0.93. Each mean correlates to the peak present in the violin graph. These results indicate that the misfolding simulation experiments had an impact on the distribution of LiP-MS ratios, however, the effect seems to be relatively small compared to the control. This is the opposite of what we hypothesized to happen. Based on the fact that both treatments are effective tools for denaturing proteins, it was hypothesized that there should be a relatively dramatic increase to protease accessibility and that the relationship would be positive, with both treatment severely increasing the protease accessibility of all the proteins in sample. As the results indicate, this did not happen. There was very little change across the proteome even though the proteins should have been thoroughly denatured before protease treatment making them much easier to digest. The difference in what we expected to see from the misfolding simulation and the actual observed values further indicates that LiP-MS has clear limitations. It is difficult to reproduce consistent results within expected ranges based on the state and structure of the protein. Without a clear pattern in the results LiP will fail to produce the necessary fidelity needed for a medical screen.

To remedy these observations of the standard protocol on CBF samples, we sought to optimize the LiP-MS protocol. First, we tried to see the effect on the limited proteolysis results if we increased the ratio of Proteinase K in solution or the digestion time with proteinase K. Ideally, if the amount of Proteinase K is increased, the areas that become accessible to Proteinase K will also increase, hypothetically allowing the LiP-MS protocol to visualize more areas of the protein structure. The violin graph's width used in figures 1, 5, 6 and 7 indicate where LiP ratios clump together in specific data sets. The peaks of the violin graphs can correspond to the mean and median of data sets. Specifically, we increased the amount of Proteinase K to a 1:25 dilution ratio in one aliquot and 1:50 in another aliquot. In Figure 7, the 1:25 and 1:50 proteinase K dilution ratio aliquots show an increased number of LiP-MS ratios observed between a range from 0 – 1.25 compared to the control aliquot with a grouping of LiP-MS ratios around 0 when compared to the standard 1:100 proteinase K dilution treatment. Using violin graphs to show the distribution LiP-MS ratios from 0 to 3 we can look at how varied LiP-MS ratios are per experimental aliquots. The places of increased width in the violin plot correspond where a large portion of data groups together. If the peak of the width is near 1, then much of the LiP-MS data groups near 1. The increase in LiP-MS ratios towards zero causes clear clumping in the violin graph as depicted by the width of the violin graph. The clumping of LiP-MS ratios in the graph indicates less variety of LiP-MS ratios. The decrease in variety leads to regions of protein that are protease accessible containing a similar LiP-MS ratio value to their neighbor. The lack of difference between LiP-MS ratios from region to region removes any meaningful analysis that can be made about individual protein regions and thus hinders any effort to build an accurate structural map from LiP-MS ratios. Even distribution of LiP-MS values allows for more accurate analysis and structural models due to many points of references available. Therefore, increasing the concentration of Proteinase K does not improve the LiP-MS protocol for cerebral spinal fluid samples.

Commented [MB1]: Alter this based on the new figures created. Explain why a vareity of LiP ratios is important.

Second, we tried to modify the amount of time that proteinase K (PK) is allowed to digest proteins in solution for LiP-MS aliquots to see if the clumping of LiP-MS ratios can be improved. We increased the amount of time that a 1:100 proteinase K reaction was allowed to digest proteins in the

solution; where one aliquot had a digestion time of 3 minutes and another allowed proteinase K to digest for 5 minutes. The violin graph in Figure 7 shows all the LiP-MS ratios from a cerebral spinal fluid sample gathered for modification of proteinase K treatment. In contrast to changing the dilution ratio of proteinase K in LiP-MS aliquot, the digestion time modifications show a more even distribution LiP-MS ratios. The more even distribution of LiP-MS ratios has the potential to provide more information on protein structure. The 3-minute proteinase K digestion time when compared to the control, digested by PK for only 1 minute, shows LiP-MS ratios spread more evenly from 0 to 1 with less clumping towards 0 or 1 (Figure 7). The 5-minute digestion has a similar trend however, the spread of the LiP-MS ratios showed greater clumping of LiP-MS ratios towards 0 and 1 compared to the 3-minute digestion. Surprisingly, the 3-minute digestion protocol change yielded the most informative spread of LiP-MS ratios. Initially it was hypothesized that increased digestion time would lead to more even distribution due to the proteins extended exposure to PK leading to more regions being targeted and digested. However, these results indicate that relationship of digestion time and more varied LiP-MS distribution is not linear.

Another experimental modification we tried was altering the LysC digestion step. LysC is an additional reagent added to LiP-MS aliquots after the initial PK digestion and before the standard trypsin digest. LysC is an endo-proteinase that cleaves at the Carboxyl end of Lysine residues. The introduction of LysC is supposed to assist in the fragmentation of peptides that occurs during the standard trypsin digest. To test if treatment with LysC could affect LiP-MS ratio distribution, we treated 3 LiP-MS aliquots with three different LysC treatments. The control for this treatment was a standard LsyC treatment with fresh LysC. The experimental aliquots had two separate treatments. One treatment involved the use of an older stock of LysC enzyme to see if the age could affect LysC activity. The second experimental modification involved omitting the LysC treatment to test whether LysC is necessary for effective LiP-MS. Figure 5 shows that all the aliquots see a fair degree of clumping of the LiP-MS ratios centered around 1. This is consistent with the mean values for each data set. The aged LysC has a mean of 1.28. The No LysC treatment has a mean of 1.17 and the control using fresh LysC has a mean of

1.01. Consistent with its mean value, the older stock of LysC and no LysC treatments show clumping of LiP-MS ratios higher than the control. In particular, the peak of the no LysC aliquot centered on 1. As mentioned before, LiP-MS ratio data between 1 and 0 is more valuable to LiP-MS analysis than LiP-MS ratio data above 1. This makes the no LysC treatment less effective than the control or use of the older stock of LysC, as a significant portion of LiP-MS data occurs over 1. The results indicate that LysC and its quality is an important LiP-MS reagent, specifically, using the standard protocol with a fresh LysC stock is the most effective LiP-MS treatment of CBF samples.

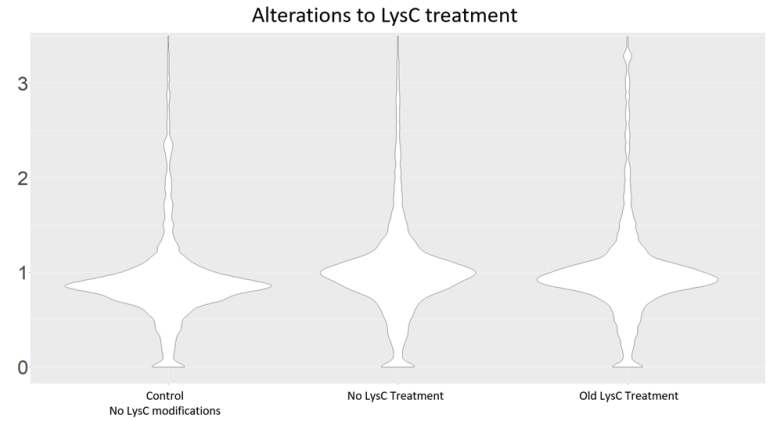

Figure 5. Mass spectroscopy analysis of misfolding simulation and reagent modification
LiP-MS ratios from every aliquot available in the limited proteolysis mass spectroscopy analysis were then separated by aliquot type to observe which experimental control produced more varied results. The violin graph shows LiP-MS ratio distribution concentrated at 1.0 across the board, with certain aliquots seeing more concentration that others. Both the no LysC and aged LysC treatments had little effect to overall LiP-MS ratio distribution compared to the control. The Sodium DOC treatment had a stronger effect on LiP-MS ratio distribution than temperature treatment.

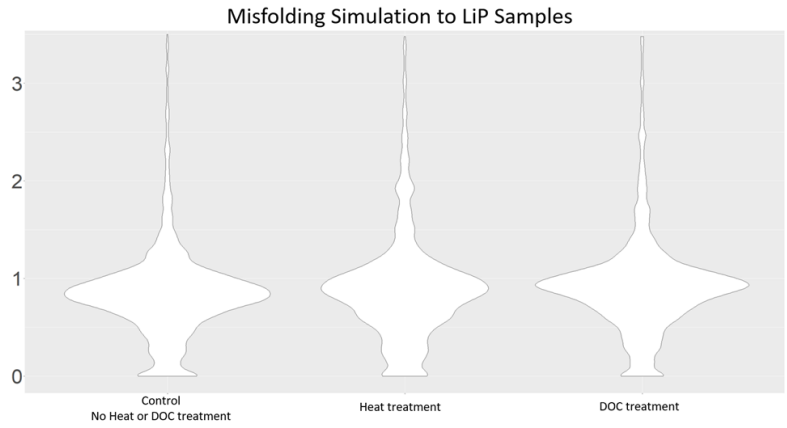

Figure 6. Mass spectroscopy analysis of misfolding simulation and reagent modification
LiP-MS ratios from every aliquot available in the limited proteolysis mass spectroscopy analysis were then separated by aliquot type to observe which experimental control produced more varied results. The violin graph shows LiP-MS ratio distribution concentrated at 1.0 for all the samples, with certain aliquots seeing more concentration that others. Both the no LysC and aged LysC treatments had little effect to overall LiP-MS ratio distribution compared to the control. The Sodium DOC treatment had a stronger effect on LiP-MS ratio distribution than temperature treatment.

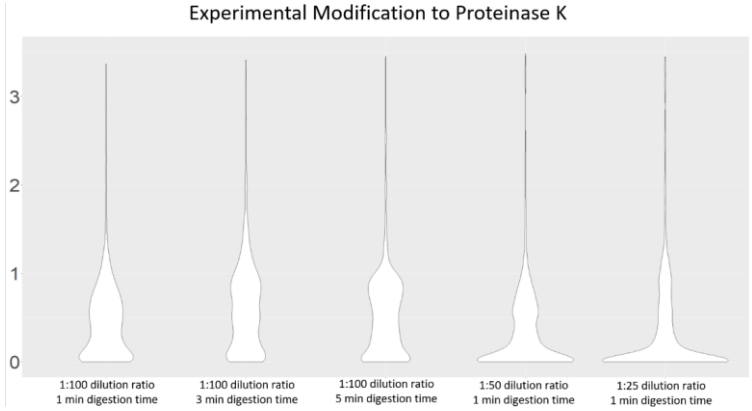

Figure 7. Mass spectroscopy analysis of dilution ratio and Digestion time experimental controls.
LiP-MS ratios from every aliquot available in the limited proteolysis mass spectroscopy analysis were then separated by aliquot type to observe which experimental control produced more varied results. The violin graph shows increased concentration of LiP-MS ratios around 0 for the 1:25 dilution ratio and 1:50

dilution ratio aliquots. The Digestion time aliquots show a similar distribution to the normal control, with more values near 1.0.

Discussion

Proteomics as a field has made major contributions to the understanding of celluar biology and developed technologies for studying biological processes as they relate to protein abundance and function. Proteins are integral to understanding phenotypic and genotypic realtionships in orgnainisms as they orchstrate all celluar processes. Therefore, a better understanding of protien dynamics allows researchers and scientist in the field of proteomics to look closer at the underlying principles of life through the lens of protein interaction in the cell. Traditionally, Proteomics focused on measuring protein dynamics by observing relative amouts of protien in a sample to explain celluar function. As technologies advanced with the advent of NMR and X-ray crystallography, protein structure became a major focus for protein studies exploring celluar processes in cells. The relationship between protein structure and function revealed limititations to studies that purely studied protein abundance as a way to explain celluar phenotypes. Though protein abundance can be enlightening when studying how the prescene and amount of protein influnces celluar function, it fails to capture a large portion of biological regulation that is not dictated by changes in protein level. Protein structure studies hope to solve this limitiation by exploring the relationship between form and funciton of proteins. These studies help fill gaps present in protien abundance studies they also come with their own limitiation. Both X-ray crystalography and NMR have size exclusions with NMR being incapable of analyzing proteins larger that 40 kDa and X-ray crystallography working well with larger protiens. X-ray crystallograpy also has limitations when looking at the amount of purifed protein analysis required for successful analyisis. X-ray crystalography also requries crystalizing the protein before analysis, which can be both time consumuing and diffcult.

Recent advancements in structure studies have lead to the use of mass spectroscopy for efficent analysis of protein sequence. Specifically, shotgun proteomics helps piece together whole protein sequcences from the peptide sequesces analyzed via mass spectroscopy. Cells with the desired protein

mixture are digested with a protease and then run through a microcapilary tube where they can be seperated by hydrophobicity and charge. The fingerprint of each peptides fragmentation is used to identify the protein they originate from by searching against a sequence database with available software. Shotgun proteomics paired with other mass spectroscopy methods can provide important information on stability and structure of proteins. Both Thermal Proteome Profiling (TPP) and Limited proteolysis, when paired with shotgun proteomics, provide information on protein stability and structural alterations. Limited proteolysis mass spectroscopy is a novel and useful method for exploring proteomics in a distinct way. Specifically, LiP-MS can provide data on protein misfolding in complex matrices, like blood or cerebrospinal fluid. Proteases acessibility offers a multiple avenues for exploring protein dynamics in a variety of solutions. Thus, LiP-MS has potential as a powerful screening tool for genetic and neurodegenrative diseases that revolve about misfolding proteins forming aggregates that hinder cell function. This thesis further explores LiP-MS optimization and its ability to analyze data on a proteome-wide scale. The development of LiP-MS as a screening tool is important for furthering the understanding how protein aggregates lead to genetic and neurodegerative diseases. In the past protein aggregation was thought as a inconvience of protein research. Recently, protein aggregation has been linked to a varitey of genetic and neurodegenerative diseases including Alzheimer's Disease, Parkison's Disease and Hunting's Disease. Amyloid fibrils are thought to be a major cause of many of the symptoms associated with alzheimer's disease. Amyloid fibrils are formed from filaments created by the spontanous aggreagation of peptide and proteins in the cell. Amyloid fibrils have a variety of peculiar characteristics including an unusual resistance to proteolytic degradation. Once the amyloid fibrils are formed they persist in the cell and can lead to cytotoxic condtions. Their stability in the face of proteases make them an ideal for LiP-MS study as their resistance to proteases degradation will produce unique LiP-MS patterns that can be relied to indicate the presence of amyloid fibirils in solution along with providing information into how certain proteins, when they become misfolded, lead to genetic and neurodegrative diseases. The optimization LiP-MS is integral to studying and combating these disease and creates opportunites to

screen for potentially misfolded proteins before the formation of amyloid fibrils in the cell and the onset of its symptoms. At the moment, LiP-MS is not accurate or reliable enough for use in medical screening. In the future, LiP-MS may be used to study more intricate proteins and how strucutral alteration can alter their interactions with other proteins. Specifcally it might find use in the study of proteins responsible for controlling genetic regulation.

Our data showes that LiP-MS protocol still needs a to altered to increase its efficiency and accuracy. Changes to the temperature and reagents used do not seem to alter the LiP ratios in a way that we predicted and alterations to the dilution ratio and digestion of Proteinase K also revealed that changes to the protocol failed to lead to predictable results. The major takeway from this research is that even more effort and time need to be put towards altering the experimental conditions of LiP-MS for increased accuracy and efficiency. If LiP-MS results patterns can be better elucidated with repeated experimetation, it will get closer to use as a complex screening tool that will help lessen neurodegenrative disease pathologies.

References

Bross, P., Corydon, T. J., Andresen, B. S., Jorgensen, M. M., Bolund, L., & Gregersen, N. (1999). Protein misfolding and degradation in genetic diseases. *Human Mutation*, *14*(3), 186–198. https://doi.org/10.1002/(SICI)1098-1004(1999)14:3<186::AID-HUMU2>3.0.CO;2-J

Fontana, A., De Laureto, P. P., Spolaore, B., Frare, E., Picotti, P., & Zambonin, M. (2004). Probing protein structure by limited proteolysis. *Acta Biochimica Polonica*, *51*(2), 299–321. https://doi.org/10.18388/abp.2004_3573

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery*, *2*(2), 140–150. https://doi.org/10.1038/nrd1011

Kaur, U., Meng, H., Lui, F., Ma, R., Ogburn, R. N., Johnson, J. H. R., Fitzgerald, M. C., & Jones, L. M. (2018). Proteome-Wide Structural Biology: An Emerging Field for the Structural Analysis of Proteins on the Proteomic Scale. *Journal of Proteome Research*, *17*(11), 3614–3627. https://doi.org/10.1021/acs.jproteome.8b00341

Mateus, A., Savitski, M. M., & Piazza, I. (2021). The rise of proteome-wide biophysics. *Molecular Systems Biology*, *17*(7). https://doi.org/10.15252/msb.202110442

Monti, M., Amoresano, A., Giorgetti, S., Bellotti, V., & Pucci, P. (2005). Limited proteolysis in the investigation of β2-microglobulin amyloidogenic and fibrillar states. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1753*(1), 44–50. https://doi.org/10.1016/j.bbapap.2005.09.004